5.7 An 8-Core 64-Thread 64b Power-Efficient SPARC SoC

Umesh Gajanan Nawathe, Mahmudul Hassan, Lynn Warriner, King Yen, Bharat Upputuri, David Greenhill, Ashok Kumar, Heechoul Park

SUN Microsystems, Sunnyvale, CA

Niagara2 is the 2nd in a series of multi-core, multi-threaded 64b SPARC processors based on the chip-multithreaded (CMT) architecture optimized for space, power, and performance (SWaP). The chip (Fig 5.7.1) has 8 SPARC cores (SPCs) that communicate with the 8-bank L2 cache through a crossbar, which provides a BW of ~400GB/s. It has one ×8 PCI-Express channel, two 10G Ethernet ports with XAUI interfaces, and four memory controllers each controlling 2 FBDIMM channels with a peak link rate of 4.8Gb/s. The FBDIMM architecture provides ~2× the memory BW at less than half the pin-count compared to DDR2. All the SerDes are on-chip, providing an aggregate BW of greater than 1Tb/s. The 8 SPCs support concurrent execution of 64 threads. Each SPC has a dedicated floating-point and graphics unit, two integer execution units and uses load/store and branch speculation to achieve high single-thread performance. The high level of system integration truly makes Niagara2 a system-on-a-chip, thus reducing system component count, complexity, and power and hence improving reliability. The chip is fabricated in an 11M 1.1V 65nm triple-V_t CMOS process and has ~500M transistors on a 342mm^2 die (Fig. 5.7.7) packaged in a flip-chip glass ceramic package with 1831 pins.

Niagara2 has a true random number generator on chip and a cryptographic unit within each SPC. They implement various encryption/decryption algorithms, hash functions, checksum generation, modular arithmetic and provide an aggregate BW of 40Gb/s, matching the combined BW of the two 10G Ethernet Ports. The cipher algorithms supported are RC4, AES, DES, and 3DES. The hash functions implemented are SHA-1, SHA-256, and MD5. This helps Niagara2 support secure applications with minimal performance cost.

One of the biggest challenges of SoC integration is the presence of several synchronous and asynchronous clock domains. Clusters are composed flat to enable better design optimization, but custom clock insertion and routing is used to maintain clock skew within tight budgets. Clock-tree synthesis is used for asynchronous clocks. Asynchronous domain crossings are handled using FIFOs. An on-chip PLL generates the ratioed synchronous clocks (RSCs) with support for a wide range of integer and fractional divide ratios. Balanced use of H-trees and grids in the clock distribution ensures low power and clock skew. The periodic relationship of the RSCs is exploited to perform high-BW skew-tolerant domain crossing with minimal reliance on accurate clock balancing. All clocks start deterministically with respect to a common reference. As shown in Fig 5.7.2, an edge-detect circuit generates a pulse 'aligned', which is delayed through the clock tree (Flops A, B) to track core clock latency. This pulse signifies that the rising edges of all the RSCs are virtually aligned at the destination cluster. It also starts a counter that increments every fast clock (FCLK) cycle until it is reset by 'aligned'. The counter value is decoded to generate exactly one 'Sync_en' pulse per slow clock (SCLK) cycle. This enables data transfer on the rising edge of FCLK in both directions (Fig. 5.7.3). The location of the 'Sync_en' pulse is chosen closest to the center of SCLK cycle. This makes it skew tolerant by equalizing setup and hold margins. In general, if 'N:M' is the FCLK to SCLK frequency ratio, 'T' is the FCLK period, and 'k' is any SCLK cycle, the margin bound for any value of 'k' is:

[0.5(N/M)T - 0.5T - skew] < t_margin < [0.5(N/M)T - skew]

PCI-Express, XAUI, and FBDIMM SerDes share a common microarchitecture. The major difference is that FBDIMM uses V_{SS}-referenced signaling versus V_{DD}-referenced signaling for the other two. A level shifter allows reuse of the NMOS-based receiver (Rx) input stage rather than using a lower-mobility PMOS-based circuit. The R_X also contains an electrical idle (EI) detector (Figs. 5.7.4, 5.7.5) to detect a quick reset, which occurs when the remote transmitter becomes active after being in EI state (Both INp and INn below 65 mV). PFETs P1 and P2 level shift 'V_{INp}' and 'V_{INn}' to 'VLS1' and 'VLS2'. NFETs N1 and N2 act like an analog 'OR' gate increasing 'V_{pk}' if either 'VLS1' or 'VLS2' rises. Capacitor C helps retain 'V_{pk}' over cycle transitions. 'V_{pk}' is then compared with the reference 'Vref' to detect EI. Bias currents 'Ib1' and 'Ib2' are generated using a circuit consisting of a bandgap reference, precision resistors, and current mirrors.

Each SPC in Niagara2 has 16KB I-cache (8-way, 32B line size) and 8KB D-cache (4-way, 16B line size). The 4MB 8-bank L2 cache is 16-way set-associative with a 64B line size. Addresses can be hashed to distribute accesses across different sets in case of hot cache sets caused by reference conflicts. Data from different ways and different words are interleaved to improve SER. Both row and column redundancy is implemented. Significant saving in X-decoder area is realized by physically locating spare rows for one array in the adjacent array, which is not required to be enabled. As shown in Fig 5.7.6, when redundancy is enabled, the incoming address is compared with the address of the defective row and if it matches, the adjacent array (which is not normally enabled) is enabled to read/write into the spare row. Memory cell nwell power is separated out as a test hook to screen out weak memory bits susceptible to read disturb fails due to PMOS NBTI effect. This improves reliability and reduces defective parts per million (DPPM).

Niagara2 uses extensive clock-gating, power-throttling, and power-down techniques for power management. To reduce leakage power, logic gates with nominal channel length are selectively replaced by footprint-compatible gates with longer channel length. Candidate gates for replacement are identified based on available slack with respect to the timing, noise, and slew specifications. Low-V_t gates are selectively used to speed up critical paths. Routing width/space combinations are chosen to optimize the power-delay product. On-chip thermal diodes monitor the die temperature and allow software to control instruction issue rates as well as turn threads on or off to manage power consumption.

The unknown phase alignment of the on-chip SerDes presents a challenge in terms of testability and debug. To aid testing, a deterministic test mode is used to eliminate the uncertainty of asynchronous domain crossings. Niagara2 employs an extensive DFM methodology that includes using certain larger-than-minimum design rules, OPC simulation on critical layouts, and extensive statistical simulation. Only analog and SRAM circuits use custom design styles. These circuits were proven on testchips prior to first silicon. Architecture design enabled use of less than 8 SPCs/L2 banks, thus shortening the debug cycle by making partially functional die usable, and providing the additional advantage of increasing overall yield by enabling partial core products. First silicon was fully functional and booted Solaris within the first week of testing.

Acknowledgements:
Sun Microsystems' Niagara2 team.
Texas Instruments for Niagara2 fabrication and co-developing SerDer.

References:

[1] Greg Grohoski, et al., "Niagara2: A Highly Threaded Server-on-a-Chip," *Hot Chips Symposium*, Aug., 2006.
[2] Ana Sonia Leon, et al., "A Power-Efficient High-Throughput 32-Thread SPARC Processor," *ISSCC Dig. Tech. Papers*, pp. 98-99, Feb., 2006.

 1-4244-0852-0/07/$25.00 ©2007 IEEE.

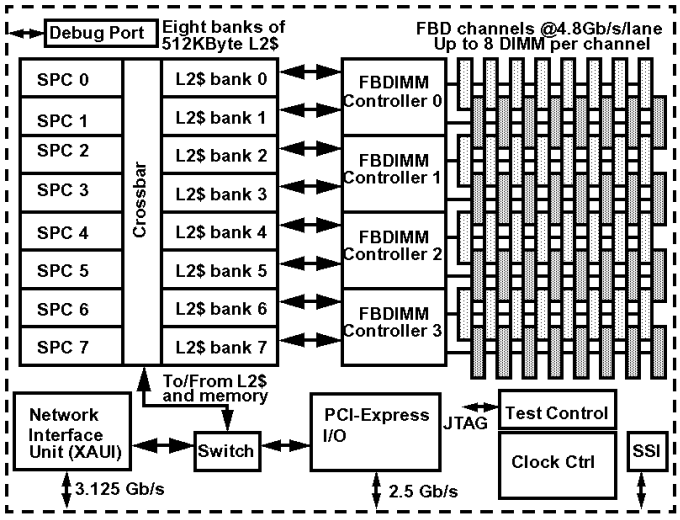

Figure 5.7.1: Niagara2 block diagram.

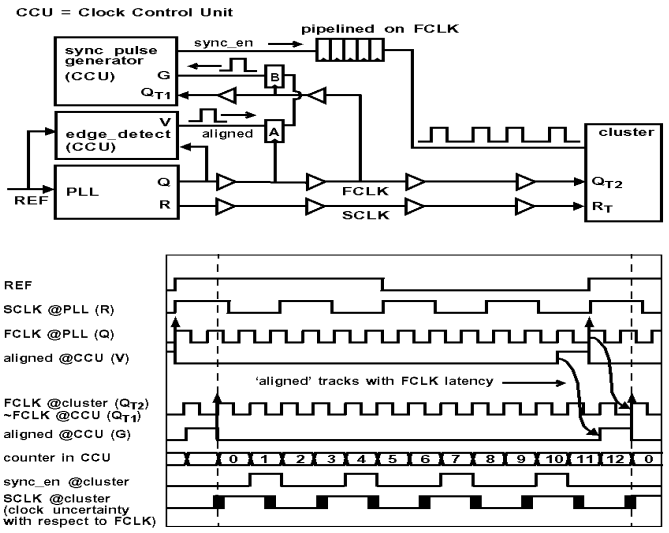

Figure 5.7.2: Generation of 'Aligned' and 'Sync-en'.

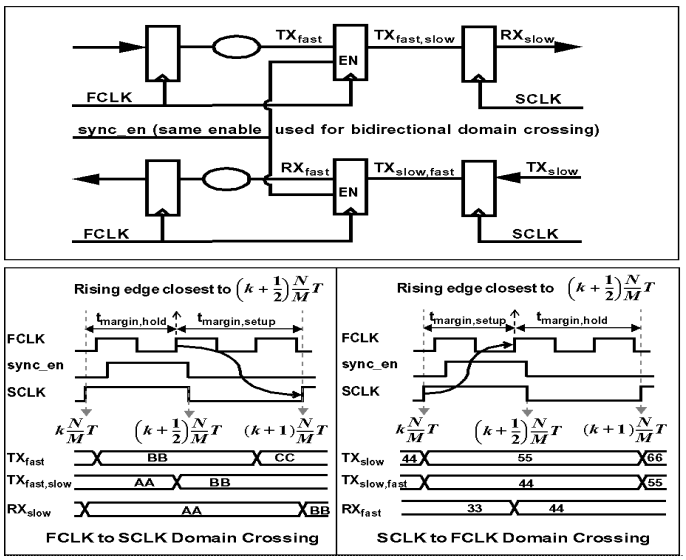

Figure 5.7.3: Ratioed synchronous clock domain signal crossings.

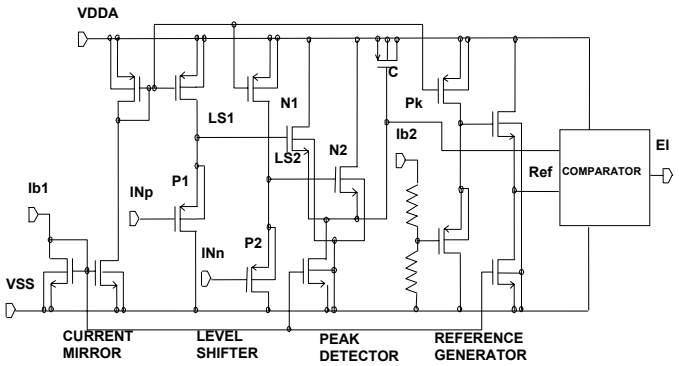

Figure 5.7.4: Electrical Idle (EI) detector.

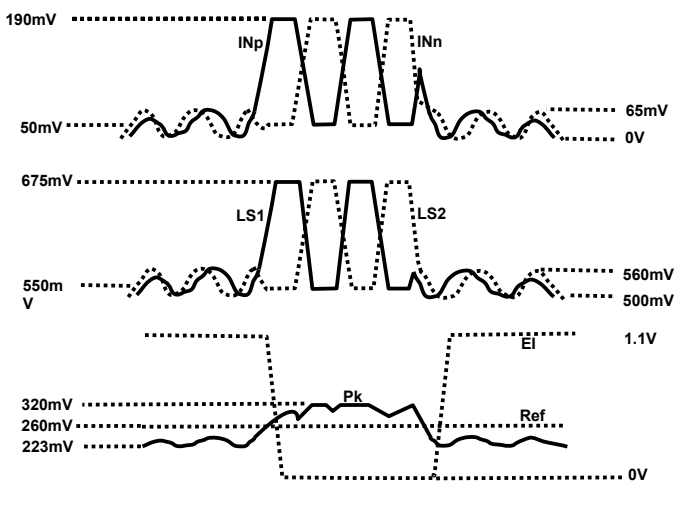

Figure 5.7.5: Electrical Idle (EI) detector waveforms.

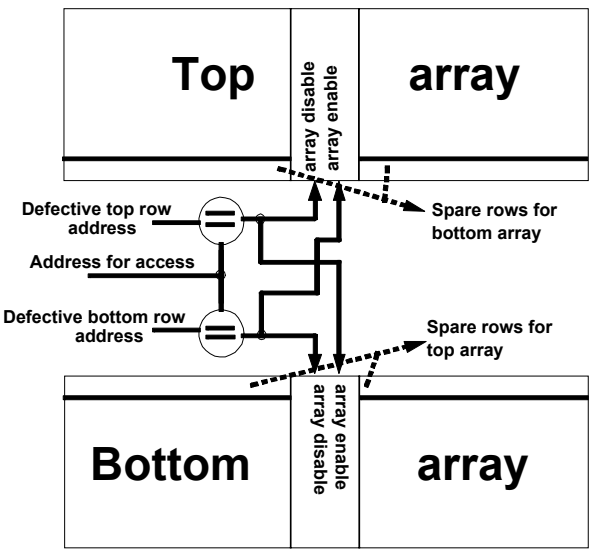

Figure 5.7.6: L2 Cache row redundancy scheme.

Continued on Page 590

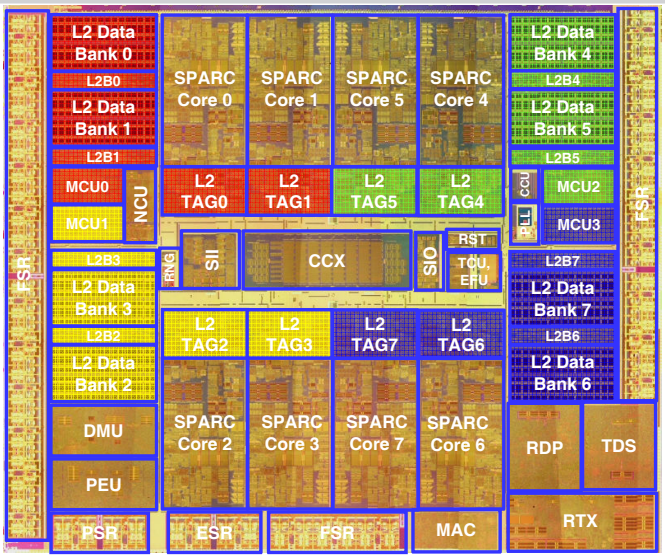

Figure 5.7.7: Niagara2 die micrograph.

 1-4244-0852-0/07/$25.00 ©2007 IEEE.